AD-A035 393 MINNESOTA UNIV MINNEAPOLIS DEPT OF PSYCHOLOGY F/G 5/10
EFFECTS OF ITEM CHARACTERISTICS ON TEST FAIRNESS.(U)
DEC 76 S M PINE, D J WEISS N00014-76-C-0244
UNCLASSIFIED RR-76-5 NL

1 OF 1
AD-A
035 393

END
DATE
FILMED
3-15-77
NTIS

EFFECTS OF ITEM CHARACTERISTICS ON TEST FAIRNESS

MINNESOTA UNIVERSITY, MINNEAPOLIS

DECEMBER 1976

EFFECTS OF ITEM CHARACTERISTICS ON TEST FAIRNESS

Steven M. Pine
and
David J. Weiss

RESEARCH REPORT 76-5
DECEMBER 1976

PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MN 55455

Prepared under contract No. N00014-76-C-0244, NR150-383
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 76-5	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Effects of Item Characteristics on Test Fairness		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steven M. Pine and David J. Weiss		8. CONTRACT OR GRANT NUMBER(s) N00014-76-C-0244
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Minnesota Minneapolis, Minnesota 55455		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.:61153N PROJ.:RR042-04 T.A.:RR042-04-01 W.U.:NR150-383
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE December 1976 13. NUMBER OF PAGES 34
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

test fairness	differential prediction	Thorndike fairness
selection fairness	test construction	item difficulties
bias	Cleary fairness	item discrimination
test bias	computer simulation	item bias
differential validity	monte carlo simulation	peaked tests
		uniform tests

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

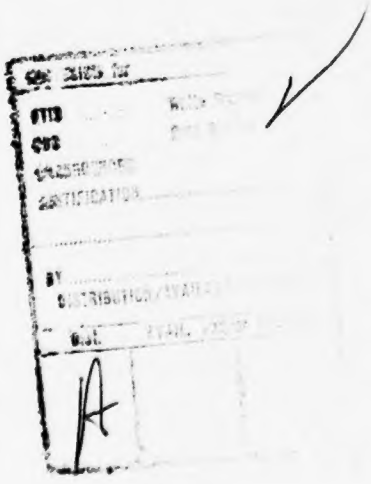
This report examines how selection fairness is influenced by the item characteristics of a selection instrument in terms of its distribution of item difficulties, level of item discrimination, and degree of item bias. Computer simulation was used in the administration of conventional ability tests to a hypothetical target population consisting of a minority and a majority subgroup. Fairness was evaluated by three indices which reflect the degree of differential validity errors in prediction (Cleary's model) and proportion of applicants exceeding a selection cutoff (Thorndike's model). Major findings

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

were: 1) tests with a uniform distribution of difficulties had fairness properties generally superior to tests having a peaked distribution of item difficulties; 2) subgroup validity differences can be expected to occur when test items are biased against one of the subgroups; 3) when differential prediction is used, the Thorndike model reflects varying degrees of unfairness due to item bias and other test characteristics, while the Cleary and validity models do not; 4) differential prediction provides fairer selection than the use of majority prediction only, regardless of the internal characteristics of the test, although substantial degrees of unfairness still exist under certain test item configurations. It was concluded that the internal characteristics of a selection instrument will affect the fairness of test scores in specific applications and that further research is needed to delineate which testing strategies and/or item characteristics are optimal in reducing unfairness.



CONTENTS

Effects of Item Characteristics on Test Fairness

Mental ability testing is commonly used in education, industry and the military services to select and place individuals. Test results are also used in research as a basis for making inferences about the intellectual endowment of various individuals and subgroups. However, many of these tests have often been cited as being biased and/or unfair to certain subgroups of the general population, including Blacks, Spanish-speaking Americans and Native Americans. Because of the prevalence of testing in our society and because of the possible discriminatory nature of some tests, there has recently been an increase in research on the nature and degree of test bias and test fairness in various settings, including examination of various ways to reduce test bias and unfairness where they exist.

A necessary prerequisite for carrying out meaningful research in this area is to define exactly what is meant by bias and unfairness. Over the last ten years, a number of models have been proposed to provide such definitions. Many of these models are quite different in philosophy and purpose. A useful taxonomy often suggested (Flaugher, 1974; McNemar, 1975; Pine & Weiss, 1976). is to separate models of bias from models of fairness. The essential distinction is that models of bias represent the psychometric properties of a particular set of test items or test scores. Models of test fairness typically are concerned with the impact a test will have when used in a particular application. The application most often considered is the selection or placement of personnel.

However, there is a direct relationship between the item characteristics of a test, including the degree of item bias, and its fairness when used in a selection program. Although substantial amounts of research have dealt with the effects of item characteristics on test validity (Brogden, 1946; Gulliksen, 1945; Tucker, 1946; Urry, 1969), no efforts have been made to study the effects of item characteristics on test fairness. Even for validity, the effects of possible bias in the test items have not been considered.

There are a number of possible reasons for this lack of research. First, selection fairness models are relatively new. Second, empirical investigation in this area is often expensive, impractical due to the relative unavailability of minority group members, and hampered by the absence of a suitable, unbiased criterion measure. Furthermore, in selection of fairness models, tests are considered only in terms of their final scores. Therefore, the internal properties of a test are generally ignored. This approach is detrimental to the development of tests which might be designed to reduce unfairness.

This report offers a general method for examining the relationship between selection and placement fairness and the characteristics of test items. This is accomplished by conceptualizing bias and fairness in terms of latent trait theory. Criterion performance is represented by the latent trait. Item bias and other item characteristics are expressed in terms of latent trait parameters. This approach eliminates the possibility that the criterion itself may be biased, and permits direct observation of how the characteristics of a test affect the prediction of a criterion and, in turn, selection fairness.

Bias and Fairness

Bias, as it is used in this report, refers to those subgroup differences in the psychometric properties of a test which occur as a result of factors extraneous to those which a test is intended to measure. For example, mean test score differences between Blacks and Whites on a vocabulary test would be considered evidence of bias if these differences reflected the influence of cultural factors. In this case, the cultural factors would be extraneous since, presumably, the test is intended to measure verbal ability.

Most of the models of bias which have been proposed (Angoff & Ford, 1973; Breland, Stocking, Pinchak, & Abrams, 1974; Echternact, 1974) have involved comparing item difficulties among subgroups. According to this approach, a test is considered biased if its items do not have roughly the same relative difficulties for all subgroups. An item within the test is said to be biased if it is relatively more difficult for a given subgroup than are most of the other test items. Other models of bias which have been proposed involve subgroup comparison of item discriminations, mean test scores, and factor loadings (*e.g.*, Angoff, 1975; Atkin, Bray, Davison, Herzberger, Humphreys, & Selzer, 1976; Jensen, 1975).

Regardless of the specific model used, the existence of bias cannot by itself be taken as *prima facie* evidence that a test is unfair. For example, a test which includes a substantial proportion of Black slang words may be unfair when used to select college freshmen, but fair when used to select social workers for employment in the Black community. Clearly then, the fairness of a test (or test item) can only be determined by examining what caused the bias and what its eventual impact will be in a specific application.

For the specific application of tests to the selection of personnel, a number of formal definitions of fairness have been developed. One of the earliest formal definitions of test fairness in selection was based on the concept of validity. This is undoubtedly due to the fact that early legal challenges to the use of tests for personnel selection questioned test validity.

Validity model of test fairness. The validity model is primarily concerned with the legitimacy of the inferences which can be made about people's ability or performance in a specific situation based on their test scores. The validity of a test is frequently determined by calculating the correlation coefficient between the test scores and scores on an appropriate criterion for a particular subgroup. Fairness of a testing procedure has been evaluated in terms of whether there is a significant difference between the validity coefficients for various subgroups on a given test. If a significant difference does exist, this would imply that the predictions made on the basis of the test scores are not as accurate for one subgroup as for another.

In a selection situation, such a difference in validity would have several adverse effects on the subgroup having the lower correlation. First, it would decrease the variance of the predicted score distribution. Assuming the selection cutoff to be above the mean of this subgroup, as it normally would be, such a decrease in variance would lower the probability that these individuals would

exceed the selection cutoff. Secondly, the lower correlation coefficient indicates that the test does not order individuals as accurately on the criterion as it would for a subgroup having a higher validity coefficient. Consequently, if selection is based on predicted criterion performance, applicants with lower average ability will be selected from the subgroup having the lower predictive validity even in cases where the subgroups have equal mean ability.

Whether or not meaningful validity differences among subgroups occur in real selection situations is an empirical issue which has received a great deal of attention recently. The weight of the evidence (Campbell, Crooks, Mahoney, & Rock, 1973; Farr, O'Leary, Pfeiffer, Goldstein, & Bartlett, 1971; Schmidt, Berner, & Hunter, 1973) seems to indicate that meaningful differences occur with very low frequency. However, a number of issues remain unresolved regarding how to statistically test for a subgroup validity difference and what to do if it is statistically significant (*e.g.*, Standards for Educational and Psychological Tests, 1974; Flaugher, 1974).

Although research still continues on differential validity as a means of evaluating test fairness, it appears that validity is a necessary but not a sufficient condition for test fairness. In recognition of this fact, a number of specific models have been proposed for defining fairness in the context of selection.

Other models of fairness. In the context of selection, test fairness is directly interpretable in terms of the number of applicants who are selected from each subgroup of testees. Test bias influences fairness to the extent that if a test is biased, it will often produce an adverse impact on the subgroup against which it is biased. This, however, will depend on how fairness is defined and on other situational variables, such as the criterion for success and selection cutoff points.

When a test is used in the selection process, it is part of a decision strategy to select or reject potentially successful individuals for one or more available openings. Operationally, this is usually achieved by setting a cutoff score on the criterion to define successful performance, determining the corresponding predictor cutoff scores, and selecting applicants with predictor scores equal to or exceeding the predictor cutoff score.

It was previously indicated that a low test validity for a given subgroup, which is equivalent to a larger amount of random errors of prediction, can affect selection decisions by decreasing the probability that individuals from that subgroup would exceed a given cutoff on the criterion. Another factor which would affect the prediction of criterion performance in selection is constant errors of prediction. The random and constant errors of prediction can be respectively translated by regression theory into the slope and intercept of the regression line relating test scores to criterion performance.

Cleary (1968) developed a widely used definition of selection fairness, referred to by her as 'bias', which involves the regression line in prediction. According to Cleary, "A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed,

consistent nonzero errors of prediction are made for members of the subgroup..." (Cleary, 1968, p. 115). Theoretically, consistent zero errors of prediction are assured by employing separate within-subgroup regression lines, *i.e.*, differential prediction. Therefore, the application of Cleary's definition is operationally equivalent to endorsing differential prediction in selection.

This fact can be demonstrated by considering the situation in Figure 1. Figure 1a illustrates the situation in which the mean criterion score for the minority subgroup ($\overline{Y}_{min}$) is equal to the mean criterion score for the majority subgroup ($\overline{Y}_{maj}$), but the mean test score of the majority subgroup ($\overline{X}_{maj}$) is greater than the score ($\overline{X}_{min}$) of the minority subgroup. In this situation it is clear that use of within-subgroup regression lines, *i.e.*, differential prediction, will produce consistent zero errors of prediction for both the minority and majority subgroups. However, using either the regression line of the majority subgroup or the regression line derived from data pooled across both subgroups will lead to underprediction of the minority subgroup.

A situation more commonly found in extant practice (Cleary, 1968; Gael, Grant, & Ritchie, 1975; Goldman & Richards, 1974; Kallingal, 1971; Temp, 1971) is where subgroups differ on both the criterion and test scores, as shown in Figure 1b. In this case, using either the majority or pooled regression line to predict minority criterion performance will result in overprediction for members of that subgroup.

In recent years, a number of models have been proposed as alternatives to Cleary's regression model of selection fairness (see Cole, 1973; and Petersen & Novick, 1974, for reviews). The one most frequently offered as an alternative to Cleary's model is Thorndike's (1971) Constant Ratio model. According to Thorndike, fair use of test scores requires that the acceptance levels should be set such that the ratio of the percentage of individuals who exceed a specified level of criterion performance to the percentage who exceed a cutoff on the predictor will be equalized among subgroups in the applicant population.

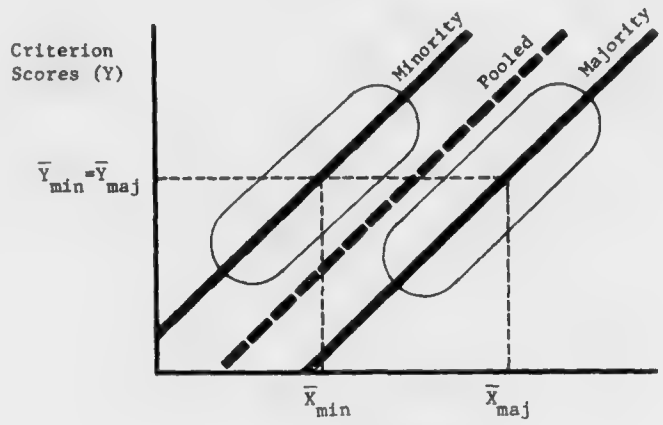
One of the primary conclusions that has derived from the research on test fairness is that the assessment of fairness will depend on how fairness is defined. Some of the models that have been proposed will lead to the selection of more minority applicants than will other models. If the models are ordered along the dimension of how many minority applicants are selected in a given situation, the Cleary and Thorndike models fall near the extremes. The Cleary model is the least favorable to minority subgroups, while the Thorndike model is one of the most favorable. Consequently, these two models make a convenient pair of strategies for evaluating the fairness of a test.

Purpose and Assumptions

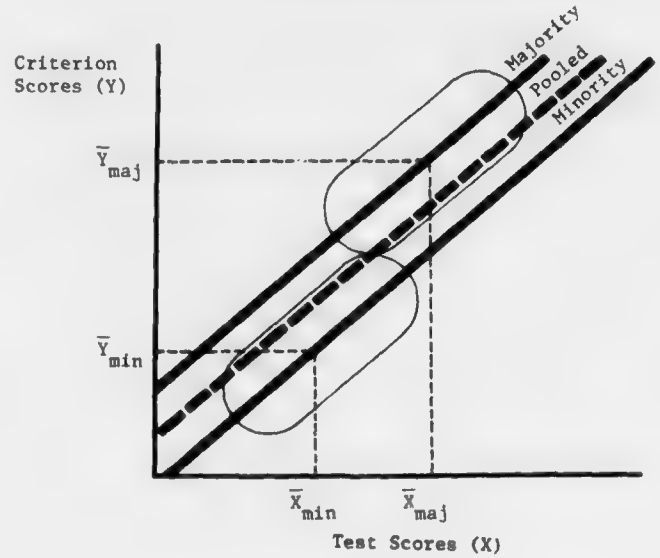
Purpose. In their book on mental test theory, Lord and Novick (1968, p. 388) indicate how the item characteristics of a test can affect the shape of the distribution of test scores. As can be seen in Figure 1, selection fairness is a function of the parameters of the distribution of test scores. Therefore, if the item characteristics of a test can affect the shape of the test score distribution, they will also influence selection fairness. The purpose of this report is to examine the relationship between characteristics of test items and selection fairness, as reflected by several fairness models.

Figure 1
Relationships between criterion scores and test scores
for majority and minority subgroups
with unequal mean scores on the predictor variables

(a). Equal Criterion Means



(b). Unequal Criterion Means



Specifically, the following questions are investigated:

1. How do the following characteristics of test items affect fairness?

   a. Distribution of item difficulties.
   b. Level of item discrimination.
   c. Degree of item bias.

2. How is fairness affected by test length?

3. How does the assessment of fairness depend on the choice of a model for measuring fairness?

Answers to these questions should be useful in indicating how a fair test should be constructed.

Assumptions. The above questions were investigated in the context of an assumed selection situation which was modeled by a monte carlo simulation study. The selection process consisted of administering a selection test to each applicant and using the score from that test to predict an external criterion represented by the known latent trait, θ. The applicant population was assumed to consist of two subgroups having identical ability distributions on θ. The selection instrument was assumed to be completely described in terms of its latent trait parameters so that each of its items could be described in terms of item discrimination, item difficulty, and probability of being guessed correctly by chance. Some of the items in the test, however, were assumed to be biased against the minority subgroup and the degree of their bias was expressed in terms of the latent-trait item parameters.

METHOD

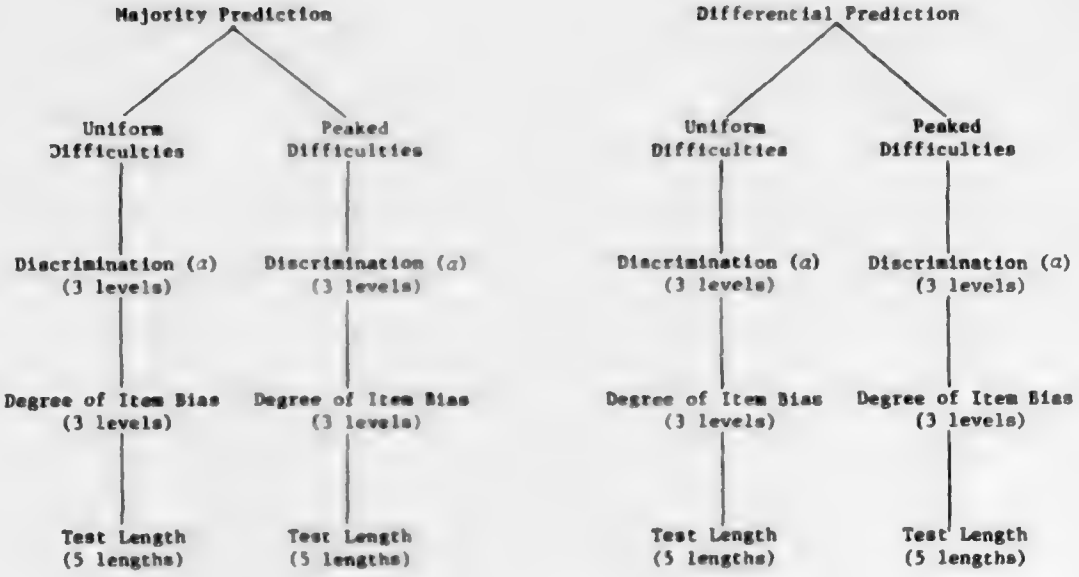
Independent Variables

Four of the independent variables were characteristic of the test administered to both the majority and minority subgroups simulated in this study. Three of these variables--distribution of item difficulties, level of item discrimination, and test length--are standard characteristics of tests. The fourth, item bias, reflected the major independent variable of interest in this study. The fifth independent variable was intended to vary the fairness in the application of test scores. This variable consisted of using only the regression equation from the majority subgroup or differential prediction, for the prediction of a simulated criterion variable. Figure 2 summarizes the independent variables used in this study.

Test Variables

Only conventional tests were used in this study. That is, all simulated testees within an experimental condition were administered identical items in a fixed sequence. Test items were represented by a set of latent trait parameters (Lord & Novick, 1968, p. 366) which described the essential statistical properties of each item. A test of length m with a given set of characteristics was generated by selecting the first m items from one of eighteen 100-item pools.

Figure 2
Independent Variables



Each item pool represented one of the experimental conditions obtained from taking combinations of the three test variables summarized in Table 1. For all experimental conditions the guessing parameter, c, was set at .20. This value is the expected proportion correct if purely random guessing occurred on five-alternative multiple-choice items.

Table 1
Item Pool Parameter Specifications

Distribution of Difficulties	a	Bias
Uniform or Peaked	.30	.5
Uniform or Peaked	.30	1.0
Uniform or Peaked	.30	2.0
Uniform or Peaked	.70	.5
Uniform or Peaked	.70	1.0
Uniform or Peaked	.70	2.0
Uniform or Peaked	1.10	.5
Uniform or Peaked	1.10	1.0
Uniform or Peaked	1.10	2.0

Distribution of item difficulties. Tests were simulated which had either peaked or uniform distributions of item difficulties. The peaked distributions of difficulties (b) were randomly sampled from a normal distribution having a mean of $\bar{b}$=0 (where 0 indicates an item of average difficulty) and a standard deviation of 1.0. The uniform distributions of difficulties also had a mean of $\bar{b}$=0 but were randomly sampled from a uniform distribution which ranged from b=-2.99 to +2.99. The actual distribution of item difficulties used in each condition is summarized in Table A in the Appendix.

Item discriminations. Three levels of item discrimination were used within both the peaked and uniform tests. These three levels were a=.30, .70 and 1.00, corresponding to point-biserial correlations of items with total scores of .127, .373 and .482, respectively (assuming a population proportion passing of P=.6 and a guessing parameter of c=.2). Values of item discrimination were held constant within each testing condition and subgroup.

Test length. To study the effects of test length and its interaction with item difficulty distributions, item discrimination, and item bias on test fairness, five typical test lengths were used. Test lengths were 10, 30, 50, 70 and 100 items. Within each test length, discriminations were constant for a given uniform or rectangular test and a specified degree of item bias.

Item bias. Item bias was defined as

$$b_{maj} - b_{min}, \qquad [1]$$

where b_{maj} and b_{min} are the latent trait difficulty parameters for the majority and minority subgroups, respectively.

This definition of item bias was based on the assumption that the subgroups had identical true ability distributions on the trait being measured, but that items were more difficult for the minority subgroup because of some independent extraneous factor(s) which reduced their performance on the test items. For example, if a test was designed to measure verbal ability, the inclusion of "culturally loaded" items would result in a test which would be more difficult for a nondominant subgroup of a given culture. The result would be a test which would be biased against such minority subgroups. This definition of item bias is very similar to those often applied in practice (Angoff & Ford, 1973; Breland *et al.*, 1974; Echternacht, 1974). The main difference is that previous models of item bias have been based on the proportion correct measure of item difficulty. However, proportion correct has been shown (Lord & Novick, 1968; Urry, 1974) to be confounded with guessing and item discrimination, whereas latent trait difficulty parameters are pure measures of item difficulty.

Three levels of item bias, based on Equation 1, were studied. These were .5, 1.0 and 2.0, indicating tests which were respectively more difficult for members of the simulated minority subgroup. Bias was introduced into the tests by adding this constant value to the difficulty parameters of the items selected to constitute the majority subgroup test. Item discrimination, guessing and test length were held constant as bias was introduced into the testing situation.

Prediction of Latent Ability

A raw test score was obtained for each simulated testee by summing the number of correct answers for that testee. Correct answers to the pth item were recorded as v_p=1, while incorrect answers were represented as v_p=0. Therefore, the raw test score for the ith individual was

$$X_i = \sum_{p=1}^{m} v_p, \qquad [2]$$

where m=test length.

Since the objective of the test was to obtain an estimate of the latent ability Θ, a method was needed to obtain a prediction of Θ based on the test score X_i. Linear regression equations were used for this purpose. Two kinds of regression equations, majority and differential prediction, were used corresponding to two types of prediction procedures often mentioned in the literature (Bartlett & O'Leary, 1969; Goldman & Hewitt, 1975; Jones, 1973; and McNemar, 1975). One regression equation of each type--majority prediction and differential prediction--was developed within each of the eighteen testing conditions. The predicted ability scores generated by these regression equations were used to define the dependent variables.

Majority prediction. In this condition, the same regression equation

$$\hat{\Theta}_i = \alpha + \beta X_i, \qquad [3]$$

where α and β are the regression parameters based on only the data from the majority subgroup, was used to predict the ability of all individuals regardless of subgroup membership.

Differential prediction. In this condition, separate within-subgroup regression equations were used to predict ability for individual i of subgroup j. These are given by

$$\hat{\Theta}_{ij} = \alpha_j + \beta_j X_{ij} \qquad [4]$$

where α_j and β_j were the within-subgroup regression parameter for subgroup j, where j referred to either the majority or minority subgroup.

Dependent Variables

The dependent variable in this study was test fairness. Fairness was evaluated by three indices separately for each of the 180 combinations of independent variables (*i.e.*, item difficulty distribution × item discrimination × test length × bias × prediction method). The three fairness indices were: 1) a validity index, R; 2) a Cleary-type index, C; and 3) a Thorndike-type index, T. These fairness measures parallel their original definitions. But in this study

the variable being predicted was θ, the known true latent ability, as compared to the fallible external criterion usually used in research on test fairness.

In addition to studying the effects of item bias and other test characteristics on these three definitions of fairness, the effects of the independent variables on a number of standard distributional statistics were also studied. These included the mean, standard deviation, standard error of estimate, skewness, and kurtosis of the ability estimates, $\hat{\theta}$.

The *R*-Index

The correlation between estimated ability and the true latent ability, $r_{\hat{\theta}\theta}$, has been used in latent trait studies as a measure of the "goodness" of ability estimation (Brogden, 1946; Urry, 1969, 1971). In the present study the true ability, θ, was taken as the criterion for selection. Therefore, $r_{\hat{\theta}\theta}$ can be interpreted as a coefficient of predictive validity. For simplicity, this coefficient of validity will be referred to simply as the *R*-Index.

Differences in *R* between the majority and minority subgroups were examined as an indication of test fairness. Larger correlations for one group as compared to the other, holding testing conditions constant, would indicate that a given set of testing conditions produced test scores with a greater potential for unfairness for the group having the lower correlation.

R was evaluated only for the majority prediction condition since the application of differential prediction amounts to a linear transformation of the majority prediction ability estimates, and correlation coefficients are unaffected by linear transformations.

The *C*-Index

Based on Cleary's (1968) concept of test bias, the degree of test bias in subgroup j can be defined as

$$C_j = \bar{\hat{\theta}}_j - \bar{\theta}_j \qquad [5]$$

where $\bar{\hat{\theta}}_j$ and $\bar{\theta}_j$ are the means of the ability distributions for the predicted and true distributions, respectively. When this definition is applied to the predicted abilities obtained from the differential prediction equation given in Equation 4, C_j=0 in all cases. This follows since $\bar{\hat{\theta}}_j$ will always equal $\bar{\theta}_j$. Consequently, the utilization of differential prediction will always result in a fair test usage according to the Cleary definition.

The inter-subgroup difference in the Cleary index is

$$C_{diff} = (\bar{\hat{\theta}}_{min} - \bar{\theta}_{min}) - (\bar{\hat{\theta}}_{maj} - \bar{\theta}_{maj}) = C_{min} - C_{maj}. \qquad [6]$$

Since in the majority prediction condition (Equation 3)

$$\bar{\theta}_{maj} = \bar{\hat{\theta}}_{maj}, \qquad [7]$$

Equation 6 simplifies to

$$C_{diff} = C_{min}. \qquad [8]$$

Similarly, in the differential prediction condition (Equation 4)

$$\bar{\theta}_{maj} = \bar{\hat{\theta}}_{maj} \quad \text{and} \quad \bar{\theta}_{min} = \bar{\hat{\theta}}_{min} \qquad [9]$$

and Equation 6 simplifies to

$$C_{diff} = 0 \qquad [10]$$

for all cases. Consequently, the Cleary index, C, was also evaluated only in the majority prediction condition.

The T-Index

Applying Thorndike's definition of fairness to the model used in this study, a test is fair if the following condition is met:

$$\frac{P(\hat{\theta}_{maj}>\theta_o)}{P(\hat{\theta}_{min}>\theta_o)} = \frac{P(\theta_{maj}>\theta_o)}{P(\theta_{min}>\theta_o)} \qquad [11]$$

where P is the proportion of testees who exceed the cutoff point θ_o. In this study a cutoff equal to the mean of the majority subgroup, *i.e.*, $\theta_o=0$ was used.

Since identical subgroup ability distributions were assumed, Equation 11 reduced to

$$\frac{P(\hat{\theta}_{maj}>\theta_o)}{P(\hat{\theta}_{min}>\theta_o)} = 1 \qquad [12]$$

or

$$P(\hat{\theta}_{maj}>\theta_o) = P(\hat{\theta}_{min}>\theta_o) \qquad [13]$$

If Equation 13 defines a fair selection situation, then the degree to which a test is unfair to the minority subgroup, as compared to the majority subgroup, is given by

$$T_{diff} = [P(\hat{\theta}_{min}>\theta_o) - P(\hat{\theta}_{maj}>\theta_o)] \times 100 \qquad [14]$$

or simply the difference between the percentage of individuals who exceed the selection cutoff in the minority and majority subgroups. The T-Index was evaluated in both the majority and differential prediction conditions.

Data Simulation

Population

The selection of examinees from a target population was simulated with a computer by generating 500 random numbers which fell between the values of -3.34 and +3.24 sampled from a normal population having a mean=0 and a S.D.=1.0. Each of the random numbers represented the true ability, θ, for one testee: θ=0 indicated an individual of average ability, while θ=3.0 indicated a person of very high ability on the relevant trait. Since the same population distribution was used for both the majority and minority subgroups, the degree of unfairness which occurred as a result of the characteristics of the test items would be manifested as differences between the predicted distributions of θ for the two subgroups. Similarly, the same 500 values of θ were used within each of the 90 experimental conditions. In this way, differences observed in the dependent variables could be attributed solely to action of the independent variables.

Simulation Procedure

The procedure used to simulate testing was carried out in three stages: 1) response vector generation, 2) application of test models to response vectors, and 3) calculation of statistics and fairness indicants.

Response generation. Generation of test responses followed procedures similar to those used by Betz & Weiss (1973), Vale & Weiss (1975) and McBride and Weiss (1976). This procedure, based on latent trait test theory (Lord & Novick, 1968), requires two assumptions. The first assumption was local independence of responses, which requires that the probability that a testee of ability θ will answer any item correctly is independent of whether that testee answers any other item correctly. Stated mathematically, this assumption becomes

$$f(v_1,v_2,v_3,\ldots,v_m|\theta) = \prod_{i=1}^{m} f_i(v_i|\theta), \qquad [15]$$

where f and f_i are probability density functions, i refers to one of the m items, and v_i=0 if a response was incorrect, and v_i=1 if correct.

The second assumption was that a response, v_i, depended only on 1) the ability of the examinee, and 2) the characteristics of the test items, as described by each item's latent trait parameters a, b and c.

With these assumptions, the response vectors were generated by:

1. Calculating $P_i(\theta)$, the probability of answering item i correctly given θ, from the normal ogive version of the latent trait test model,

$$P_i(\Theta) = c_{ij} + (1 - c_{ij}) \cdot \int_{-\infty}^{L_i(\Theta)} \phi(t)dt \qquad [16]$$

where $L_i(\Theta) = a_{ij}(\Theta - b_{ij})$,

$\phi(t)$ is the normal density function,

j indicates subgroup membership (majority or minority), and

c_{ij}=.20.

2. Determining the response v_i by:

   a. Generating a random number drawn from a uniform distribution, r, $0<r<1$.

   b. If $r>P_i(\Theta)$, $v_i=0$.

   c. If $r\leq P_i(\Theta)$, $v_i=1$.

3. Repeating this process for each item used, and for each subgroup. Two vectors of item responses were generated for each ability level for each item pool, one for the minority subgroup and one for the majority subgroup.

Test administration. The response vectors served as input to a program which simulated the testing process. Since only conventional tests were simulated in this study, the program selected items sequentially from one of the eighteen combinations of item parameters. This process was repeated for each of the five test lengths within each combination of the other sets of item parameters. Varying test lengths were obtained by selecting the first m items out of the 100 items available, where m was the desired test length (10, 30, 50, 70 or 100 items).

Application of fairness models. The output of the second stage of the simulation was an estimated Θ, $\hat{\Theta}$, for each examinee for each test condition. Therefore, a distribution of true and estimated Θ values was produced for each subgroup for each of the 90 experimental conditions. Within each of these test conditions, the mean, standard deviation, skewness and kurtosis were computed for the $\hat{\Theta}$ variable, and the validity, Cleary, and Thorndike measures of fairness (*i.e.*, R, C, T) were calculated.

RESULTS

Distributions of Predicted Scores

Means, standard deviations, skewness, and kurtosis indices of ability estimates as a function of the experimental conditions are given for a test length of 50 items in Table 2; results for test lengths of 10, 30, 70 and 100 items,

which generally parallel those for 50 items, are given in Appendix Tables B through E. In these tables, the statistics for the true ability distribution (θ) are given in the first row of the table, listed under the "True" group heading. In the standard deviation column, values obtained when differential prediction (D.P.) was used are given as well as values for the majority prediction (M.P.) case. Since differential prediction did not affect any of the other statistics, only one set of values is shown.

As Table 2 shows, increasing item bias caused the mean of the minority subgroup to be underpredicted. The degree of underprediction increased both with increasing item bias and with increasing item discrimination. For low item discrimination, the degree of underprediction was less than the degree of item bias introduced, with the degree of underprediction being somewhat larger for the peaked test at each of the item bias levels. At high item discrimination (a=1.1), the degree of underprediction became essentially equal to the degree of bias at the .5 and 1.0 levels of item bias. With item bias equal to 2.0, the degrees of underprediction (-1.85 and -1.52, for the uniform and peaked tests, respectively) more closely approached the degree of bias than did the degrees of underprediction (-1.34 and -1.10) in the low item discrimination condition at this same bias level. Also, at the high item discrimination level, the degree of underprediction was somewhat smaller for the peaked test at each item bias level.

Table 2
Score Distribution Characteristics for Conventional Tests of Length 50, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Mean		Standard Deviation				Skewness		Kurtosis	
					Uniform		Peaked					
a	Bias	Group	Uniform	Peaked	M.P.	D.P.	M.P.	D.P.	Uniform	Peaked	Uniform	Peaked
		True	-.074	-.074	1.006	1.006	1.006	1.006	-.01	-.01	.22	.22
.30												
		maj	-.074	-.074	.798	.798	.807	.807	.03	-.11	.00	-.06
	.5	min	-.391	-.402	.822	.805	.811	.820	.04	.01	-.09	-.00
	1	min	-.709	-.738	.819	.810	.824	.822	.08	.10	-.16	-.02
	2	min	-1.336	-1.097	.819	.815	.800	.816	.28	.33	.10	.17
.70												
		maj	-.074	-.074	.940	.940	.946	.946	-.08	-.11	-.27	-.66
	.5	min	-.496	-.535	.938	.939	.942	.949	.10	.19	-.33	-.66
	1	min	-.953	-.894	.929	.934	.903	.941	.31	.49	-.19	-.38
	2	min	-1.749	-1.708	.823	.920	.719	.896	.57	1.13	.08	1.08
1.1												
		maj	-.074	-.074	.967	.967	.959	.959	-.03	-.10	-.25	-1.13
	.5	min	-.536	-.536	.978	.965	.937	.957	.20	.36	-.24	-.93
	1	min	-1.020	-.956	.948	.960	.845	.937	.30	.86	-.33	-.07
	2	min	-1.852	-1.573	.813	.939	.540	.835	.77	2.13	.42	4.86

Note. M.P. is majority prediction equation; D.P. is differential prediction equation.

The standard deviation of the ability distribution was generally underpredicted, using majority prediction, both for the majority and minority subgroups. For the uniform test, the degree of underprediction was reduced for both groups as item discrimination increased, while for the peaked test, underprediction increased for the minority subgroup while it decreased for the majority subgroup.

Within the peaked test, the degree of underprediction of the standard deviations became especially severe with increasing item bias, at high discriminations.

When differential prediction was used, however, the degree of underprediction of the standard deviation was substantially reduced. Even at a=1.1 for the peaked test, underprediction of the standard deviation for the minority subgroup was virtually the same as for the majority subgroup, except for very high (2.0) levels of item bias.

The skewness for both the uniform and peaked tests increased in a positive direction as both item bias and item discrimination increased. This effect was much larger for the peaked test. At a=1.1 and bias of 2.0, the peaked test had a skewness of 2.13 compared to .77 for the uniform test. The kurtosis measure indicated that the shape of the distribution changed from being somewhat flat (negative value) to being peaked (positive value) as item bias was increased; the degree of this change was a function of increasing item discrimination. Again, the uniform test, when compared to the peaked test, more closely maintained its resemblance to the true normal distribution as bias was increased.

Fairness

Validity: *R*-Index

Effects on majority subgroup. The validity coefficients for the uniform (U) and peaked (P) distributions of item difficulties are shown in Table 3. The three rows in Table 3 labeled "maj" give the validities for the majority subgroup for the three values of item discrimination. These results correspond to the case where item bias is zero.

Validity was found to increase as item discrimination and test length increased for both types of item distributions. At the lower discrimination levels, a=.30 and .70, the peaked distribution gave higher values of validity; but at the high discrimination level, a=1.1, and for test length longer than about 40, the advantage reversed and the uniform distribution gave higher validities. The highest validity found was R=.981 for the uniform distribution of item difficulties at a=1.1, for a test length of 100 items. The validity for peaked tests at this same point was R=.967. The lowest validity also occurred for the uniform distribution. At a=.30 for test length=10, R=.493, while R=.540 for the peaked distribution.

Validity differences. A major concern with respect to test fairness refers not only to how validity varies as a function of the test characteristics for a given subgroup, but more importantly, how validity varies differentially among subgroups. The reason for this is that if a difference in subgroup validities does exist, this would imply that the predictions made on the basis of the test scores are not as accurate for one subgroup as for the other. As was explained in the introduction, such a difference in validity would have several adverse effects on the subgroup having the lower correlation. Therefore, the effects of item bias on validity were studied by comparing the validities for both subgroups for all the item pools and test lengths. To facilitate this analysis,

Table 3

Validity Coefficients for Uniform (U) and Peaked (P) Conventional Tests at Five Test Lengths as a Function of Item Discrimination (a) and Item Bias, for Majority Group (maj) and for Minority Group (min), and Differences in Validities (diff) for the two groups

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	.493	.540	.725	.741	.793	.802	.846	.848	.884	.888
	.5	min	.492	.543	.741	.754	.800	.814	.853	.860	.887	.896
		diff	-.001	.003	.016	.013	.008	.013	.007	.013	.003	.008
	1.0	min	.512	.554	.743	.763	.805	.817	.855	.860	.888	.893
		diff	.019	.014	.018	.022	.012	.016	.009	.012	.004	.004
	2.0	min	.523	.540	.749	.759	.810	.811	.855	.855	.886	.889
		diff	.030	-.001	.024	.019	.017	.009	.009	.007	.002	.001
.70												
	0.0	maj	.745	.783	.899	.912	.935	.941	.954	.955	.966	.966
	.5	min	.744	.797	.898	.918	.934	.943	.953	.956	.965	.967
		diff	-.001	.014	.000	.006	-.001	.002	-.001	.001	-.001	.001
	1.0	min	.764	.801	.891	.918	.928	.936	.949	.949	.963	.959
		diff	.019	.018	-.007	.006	-.006	-.005	-.005	-.006	-.003	-.007
	2.0	min	.773	.756	.880	.861	.915	.891	.932	.911	.950	.925
		diff	.027	-.026	-.019	-.051	-.020	-.050	-.022	-.044	-.016	-.042
1.1												
	0.0	maj	.820	.869	.932	.940	.961	.954	.973	.961	.981	.967
	.5	min	.829	.880	.937	.941	.959	.951	.972	.957	.979	.963
		diff	.009	.011	.004	.001	-.002	-.002	-.002	-.004	-.002	-.004
	1.0	min	.844	.853	.932	.921	.954	.931	.966	.937	.976	.942
		diff	.024	-.016	-.001	-.019	-.007	-.022	-.007	-.024	-.005	-.025
	2.0	min	.824	.753	.915	.818	.934	.831	.950	.837	.960	.841
		diff	.004	-.115	-.017	-.122	-.028	-.123	-.024	-.124	-.021	-.125

differences between subgroup validities were determined. Differential validity was thus defined as

$$R_{diff} = R_{min} - R_{maj}. \qquad [17]$$

A negative value of differential validity indicates that the majority subgroup had a larger validity coefficient than the minority subgroup. These values appear in Table 3 in the rows designated "diff".

Table 3 shows that for the lowest a-value, validity differences were very small for low levels of item bias. As item bias increased, differential validity increased for the uniform test, but decreased for the peaked test, except at 100 items where differential validity decreased for both tests. At a=.30, differential validity tended to be positive in favor of the minority subgroup for both types of tests. But for item discriminations of a=.7 and 1.1 for test length of 30 and above, the direction of differential validity was reversed and the tests became unfair to the minority subgroup. As the degree of item bias and item discrimination increased, the size of this negative differential became substantial, particularly for the peaked test. This effect was present at all test lengths above 10 items. For example, the peaked test at a=1.1, test length=100, and bias=2.0, had a .125 difference between the subgroup validities, in favor of the majority subgroup. The largest negative differential validity for the uniform tests was R_{diff}=-.028 which occurred at a=1.1, test length=50, bias=2.0.

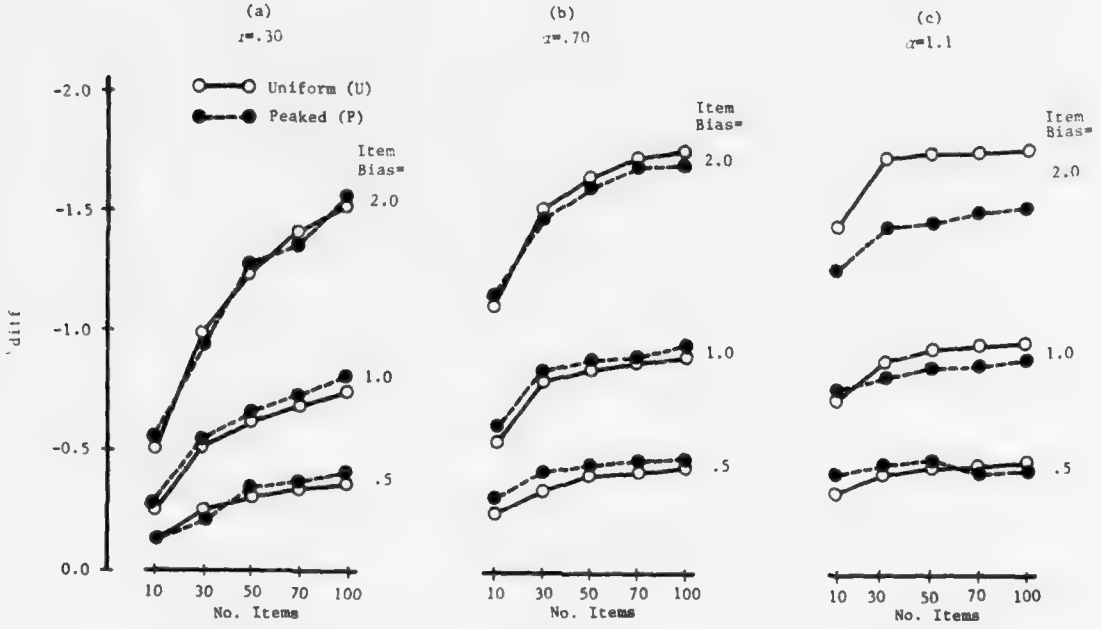
C-Index

The Cleary-type fairness measure, C, was defined as the difference between the means of the true ability, $\overline{\theta}$, and the predicted ability, $\overline{\hat{\theta}}$. Therefore, the C-Index is in the same units as θ. The population distribution of θ had a mean of 0 and a standard deviation of 1.0. A negative C-Index implies unfairness for a subgroup. In Figure 3, C_{diff}, the subgroup differences in the C-indices ($C_{maj} - C_{min}$), are plotted against test length for both the uniform and peaked tests for all item pools in the majority prediction condition. As indicated by Equation 8, under the assumptions of the present study, $C_{diff}=C_{min}$, since $C_{maj}=0$. Numerical values of C by subgroup are shown in Appendix Table F. C_{diff} was not computed under differential prediction since, as indicated earlier, by definition it is always equal to zero in this condition.

As would be expected, the C-Index indicated increased unfairness for the minority subgroup as item bias was increased from .5 to 2.0. Unfairness also increased as a negatively accelerating function of test length reaching its highest value at a test length of 100. The rate of increase as well as the highest value varied as a function of item discrimination and degree of item bias. For both the uniform and peaked tests, increasing item bias tended to increase the rate of increase of C with test length within a level of item discrimination. The effect of test length decreased as item discrimination increased.

There appeared to be very little difference between the peaked and uniform distribution of difficulties on C_{diff} at the a=.3 and .7 levels of item discrimination. The differences which do occur appear to favor the uniform tests at the

Figure 3
C-Index as a function of item discrimination (a), item bias, and test length, using majority prediction



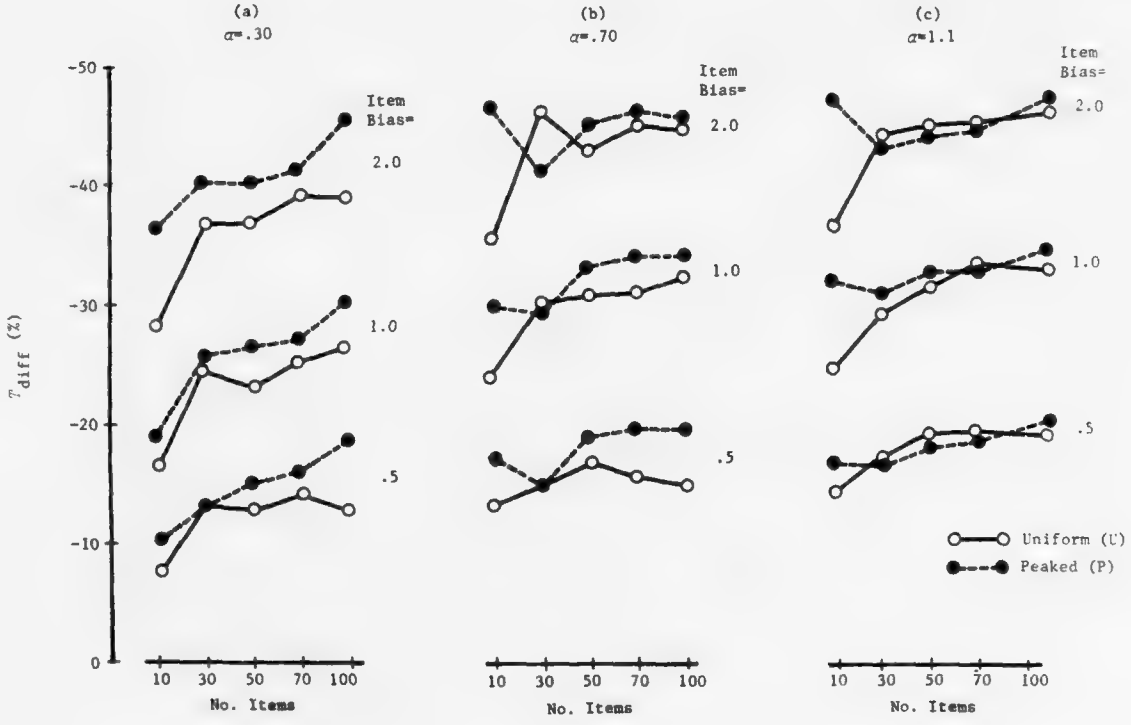
.5 and 1.0 bias levels at a=.3 (Figure 3a) and .7 (Figure 3b), and the peaked tests at bias of 2.0 when a=.7. However, at the highest discrimination level, a=1.1 (Figure 3c), the uniform tests were more unfair than the peaked tests to the minority subgroup when the degree of item bias was large (2.0). For an item bias of 2.0, differences of .350 and .342 were found between the subgroup C values for test lengths of 70 and 100 items. Thus for this test situation, using peaked instead of uniform distributions of difficulty would produce an average estimate of ability with a decrease in item bias of more than one-third of a standard deviation relative to the population of true abilities.

T-Index

T_{diff} can be defined as the difference between the T-indices for the majority and minority subgroups, *i.e.*, $T_{maj} - T_{min}$. A negative T_{diff} indicated that the percent predicted to be above average was smaller for the minority than for the majority subgroup; *i.e.*, the test was less fair to the minority subgroup.

Majority prediction. As Figure 4 shows, using majority prediction, T_{diff} varied in a complex way as a function of item discrimination, test length and degree of item bias, for the uniform and peaked tests (numerical values are

Figure 4
T-Index as a function of item discrimination (α), item bias, and test length, using majority prediction



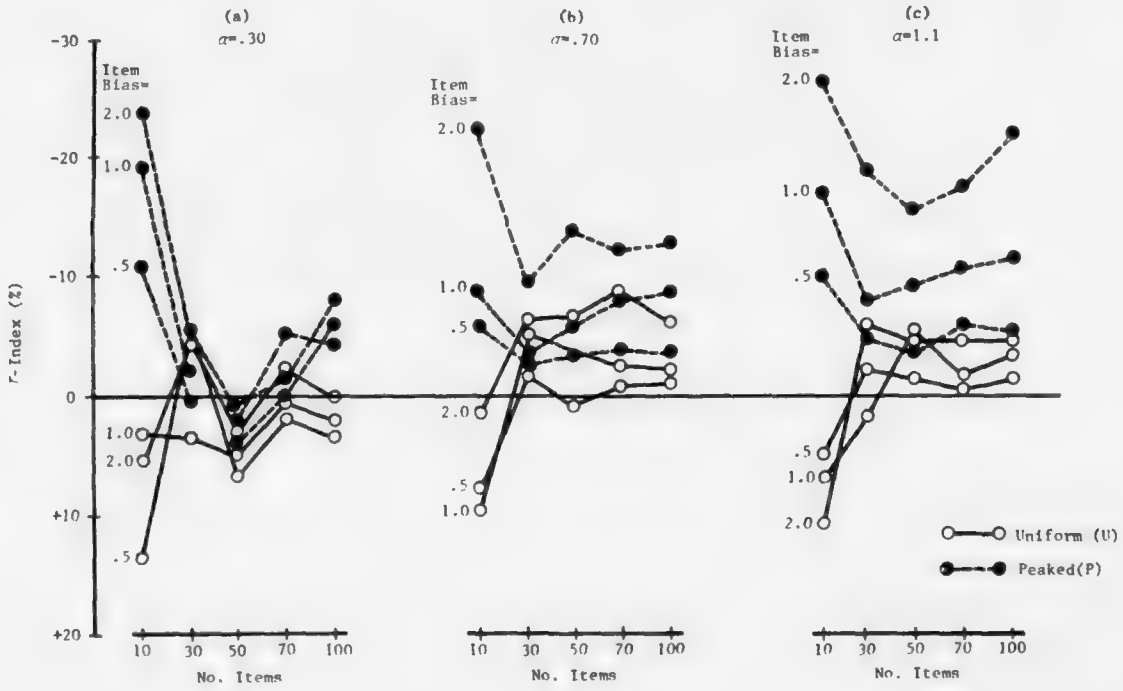
given in Appendix Table G). In general, however, the uniform tests were less unfair to the minority subgroup at the α=.3 and .7 levels of item discrimination (Figure 4a and 4b, respectively), but showed no clear advantage at the 1.1 level (Figure 4c) except for the shortest test length. Regardless of item discrimination or degree of item bias, the shortest and longest test lengths of the uniform test resulted in relatively greater fairness. Only for the intermediate test lengths did the peaked test sometimes produce a smaller T_{diff} than did the uniform test and then usually at the higher discrimination levels. In contrast to the *C*-Index, unfairness measured by T_{diff} did not increase as a regular function of test length for the peaked test at item discrimination levels above α=.30.

The largest difference in C_{diff} between uniform and peaked tests was 11.2%, occurring at the highest bias and discrimination levels at test length=10

(Figure 4c). Even for the a=.70, bias=.5, test length=100 (Figure 4b), a test which might be representative of one used in real selection situations, the uniform test would have led to the selection of 12.8% fewer minority applicants.

Differential prediction. The results of using differential prediction on T-fairness are shown in Figure 5; numerical values are in Appendix Table H. Since T_{diff} for the differential prediction case used the same T value for the majority subgroup, but a different value for the minority subgroup, results from the two prediction situations directly show the reduction in unfairness due to differential prediction. A comparison of Figure 4 with Figure 5 thus shows that

Figure 5
T-Index as a function of item discrimination (a), item bias, and test length, using differential prediction



the main effect of using differential prediction was that a much larger percentage of minority applicants was predicted above average than was the case when majority prediction was used. Consequently, the general level of unfairness was reduced using differential prediction.

Figure 5 shows that with differential prediction, the minority subgroup sometimes had a greater percentage of examinees above the mean than did the majority subgroup. This is a situation which never occurred in the majority prediction case (Figure 4). For the most part, this overprediction for the minority subgroup occurred almost entirely for the uniform test and tended to decrease as test length and item discrimination increased. Overprediction virtually disappeared for test lengths greater than 30, at item discriminations of a=.70 and 1.1 (Figures 5b and 5c, respectively). On the average, both the peaked and uniform tests tended to give higher negative values of T_{diff} as item discrimination increased, indicating increased unfairness, even using differential prediction. This effect was particularly pronounced for the peaked tests; the unfairness of uniform tests was less affected by increasing item discrimination.

The uniform tests, with only one exception, produced values of T_{diff} that were less negatively biased than the peaked tests. This superiority of the uniform tests increased as the degree of both bias and item discrimination increased. The difference was particularly large for tests of shorter length. For a=1.1, bias=2.0 and test length=10, there was a difference between the uniform and peaked tests of 23.4% in the percentages of minority testees predicted to be above average.

DISCUSSION

Effects of Item Characteristics on Validity

There has been considerable previous research (Brogden, 1946; Cronbach & Warrington, 1952; Gulliksen, 1945; Lord, 1952; Tucker, 1946) on the relationship between item statistics and test validity. It generally has been shown that the best distribution of item difficulties for maximizing validity, *i.e.*, correlation with underlying true ability, depends on a number of factors including the level of item discrimination. However, other things being equal (*e.g.*, the ability distribution peaked near the difficulty level of the items), a higher validity will be achieved with a peaked distribution of item difficulties than with a uniform distribution of item difficulties, unless items with very high discriminations are employed. This result has led many test constructors to recommend the general use of peaked tests, since the level of item discrimination at which the uniform test gives higher validity was generally thought to be too high to occur in realistic testing situations.

However, most of the previous research was conducted using conventional item statistics. It has been shown (Lord, 1975; Urry, 1974) that conventional item statistics confound the effects of guessing with item difficulty. When guessing effects are properly accounted for by using latent trait parameters, the level of item discrimination at which the uniform test produces higher validity is well within the range which occurs in common practice. This result was first reported by Urry (1969, p. 140; 1974) and was reaffirmed in the present study.

At discrimination levels of a=.3 and .7 corresponding to point-biserial correlations of item response and total score of .187 and .373, respectively, the peaked test produced a higher validity, although its advantage over the uniform test tended to decrease with increasing test length. These results are

similar to what has been reported with conventional item statistics. However, at the a=1.1 level of discrimination (corresponding to point-biserials of .48) and for tests of 50 items or more, the uniform test produced higher validities than the peaked test. For a 100-item test at a=1.1, validity was .981 for the uniform test compared to .967 for the peaked test. This represents a substantial increase in validity at this high level of correlation. Therefore, it would appear that the uniform test might be preferable in many practical situations.

Effects of item bias. When test items were biased against the minority subgroup, validity generally decreased as item bias increased (except at low item discrimination levels) for both peaked and uniform tests. This effect produced validity differences between the minority and majority subgroups since items were unbiased relative to the majority subgroup. Furthermore, these validity differences increased at a given level of item bias as item discrimination increased. The implication of these results is that if items are biased, increasing item discrimination can decrease test fairness as reflected by subgroup validity differences.

Different types of tests produced different levels of unfairness as measured by the validity index. Where item discrimination was at least a=.7, the uniform test was clearly superior to the peaked test in producing a fair test. The advantage of using the uniform test increased with increasing item discrimination and test length. With a peaked test, at a=1.1 and a test length of 100, the minority subgroup had a validity .125 below that of the majority subgroup. Under these conditions, there was only a .021 difference in subgroup validities when a uniform test was used.

These results have several implications for the construction of tests and for the interpretation of existing test data. First, they offer a possible explanation for the often-reported but controversial phenomenon of differential validity. Several researchers (Campbell *et al.*, 1973; Farr *et al.*, 1971; Schmidt, Berner, & Hunter, 1973) have presented arguments, based on various analyses of empirical data, that differential validity does not exist as a substantive phenomenon. The results of this study indicate that differential validity is a definite possibility and, in fact, can be expected when test items are biased against one of the subgroups being tested. The fact that validity differences are not often detected in practice may be due to the problem of generating sufficient statistical power to detect a difference when it exists (Bartlett, Bobko, & Pine, in press).

Thus, if test items are biased, differential validity is the expected result. Furthermore, the usual practice of selecting items having the highest item discriminations will have the effect of increasing subgroup validity differences, particularly in peaked tests.

Other Models of Selection Fairness

In the context of this study, the C-Index, based on Cleary's fairness model, gave the degree of statistical bias in the estimation of a known criterion value. The T-Index, based on Thorndike's definition of fairness, reflected the impact of estimator bias on the percentage of applicants predicted to exceed some qualifying point of ability, in this case, the mean of the population.

The Cleary view of fairness tends to optimize selection from the vantage point of the selecting institution since it assures that the ablest candidates

will be selected. The Thorndike model tends to be more liberal from the viewpoint of the minority subgroup. Even in situations where the Cleary index indicates a perfectly fair test, it has been previously shown by Schmidt & Hunter (1974) that the Thorndike index may still indicate unfairness. This result was replicated in the present study.

Furthermore, both models indicated that the nature of a test, in terms of its spread of item difficulties, can have a strong effect on fairness at some levels of item discrimination and for some test lengths. For the levels of discrimination and test lengths most commonly found in practice, the general finding was that the peaked test was fairer in terms of the *C*-Index, while the uniform test was fairer in terms of the *T*-Index, when majority prediction was employed.

The differential prediction condition indicated the conservative nature of the *C*-Index. By definition, in this condition, all tests were perfectly fair by the Cleary model. Yet the *T*-Index indicated the presence of substantial unfairness, particularly for very short tests and for highly discriminating tests. Furthermore, with differential prediction of ability, the uniform distribution of item difficulties predicted more minority testees to be above average across nearly all conditions than did the peaked distribution of item difficulties.

C-Index. One of the major trends in the data is shown in Figure 3; for both the peaked and uniform tests, the effect of item bias on the *C*-Index increased with test length. This implies that the shorter a test is, the more fair it will be in terms of producing a smaller underprediction of the minority ability level. In other words, shorter tests are less sensitive (more robust) to the presence of item bias than are longer tests. Unfortunately, this finding runs contrary both to conventional wisdom and to the results from the validity index which indicated an increase in validity with increasing test length.

The reason for this seemingly paradoxical result is that the longer a test is, the more chance there is for bias to affect the final test score. For example, if a test is only one item long, the only possible test scores are 0 and 1. Therefore, there is not as much opportunity for bias to affect the test score. On the other hand, if a test is very long, even a small degree of bias can be reflected in the score.

The influence of test length on fairness as measured by the *C*-Index was reduced, however, by increasing the level of item discrimination. What this implies is that the length of a test plays a much larger role in the ultimate fairness of a test at the lower levels of discrimination than it does at the higher levels. For example, Figure 3 indicates that if item bias is relatively large (2.0), the extent to which the minority subgroup is underpredicted will vary from 1 to 1.5 standard deviations as test length increases from 30 to 100 items. At the highest level of discrimination, however, the increase in underprediction is relatively constant between these test lengths. Consequently, in order to achieve a high level of validity and the smallest possible underprediction of the minority subgroup, the highest possible level of item discrimination should be maintained, particularly for short tests.

If a test uses highly discriminating items, the distribution of item difficulties will become an important factor in test fairness as measured by the

C-Index. For highly discriminating items, if there is reason to suspect a relatively high degree of item bias, the results of this study indicate that a peaked test is to be preferred over a uniform test. Unfortunately, this conclusion conflicts with the findings based on the validity data where it was found that a uniform test produced the smallest difference in validities with highly discriminating items. Apparently, a decision must be made as to which criterion is most important in a given situation--reduction in the difference between subgroup validities, or reduction in the underprediction of the minority subgroup.

In making this decision, the test constructor must carefully consider the degree of precision which must be sacrificed in order to reduce the relative degree of unfairness to a minority subgroup. Some minimum degree of precision must surely be maintained or one could end up with a perfectly fair, but totally useless selection instrument. This situation would, for example, be approached by employing very short tests using items with very low discrimination.

T-Index. As was the case with the *C*-Index, increasing average item discrimination had the overall effect of increasing unfairness as measured by the *T*-Index. The relationship between fairness as measured by the *T*-Index and test length, however, was more complicated than it was when fairness was measured by the *C*-Index. For some levels of item discrimination, *T*-fairness increased with test length, while in other cases it decreased. In general, however, the fairest tests were the shortest tests using the least discriminating items. This is the same result found for the *C*-Index and was, again, probably due to the restriction in the number of unique scores possible and the increased unreliability characteristic of a short test.

Results for the *T*-Index indicated that the uniform test was consistently less adversely affected by item bias than was the peaked test for the lower levels of item discrimination. However, at higher item discriminations, neither test design was obviously favorable.

Implications. Some generalizations about test design can be made based on these results. Specifically, at moderate levels of item discrimination and test lengths above 50 items, uniform tests are clearly superior to peaked tests in terms of reducing unfairness. This conclusion holds for all three fairness indices. At high item discrimination levels (above a=1.1), where uniform and peaked tests produced conflicting results in terms of validity and *C*-fairness, the distinction between distribution of item difficulties and fairness is less clear. At these levels, the distribution of item difficulties does not seem to make much difference as long as the tests are at least moderately long (greater than 30 items). Also, at these high levels of item discrimination, the expected loss in relative test validity for the minority subgroup would be small. Therefore, in view of the superiority of peaked tests in terms of *C*-fairness under these conditions, they would generally be preferable to the uniform tests.

Differential Prediction

When differential prediction is used, a test will always be fair in terms of Cleary's definition of fairness. That is, there will be no overprediction or

underprediction of mean ability level for that subgroup. Similarly, within the model used in this study, the use of differential prediction will not be reflected in the *R*-Index since it amounts to adding a constant to the scores of the minority subgroup. Such a constant will not change the correlation of test scores with another variable.

However, a test may be unfair according to the Thorndike definition of unfairness in the differential prediction condition. The degree of unfairness will depend on the item discrimination level, test length and distribution of item difficulties. As was the case for *C*-fairness and for *T*-fairness using majority prediction, differential prediction was accompanied by an overall decrease in fairness to the minority subgroup as average item discriminations increased. The relationship between *T*-fairness and test length, however, was much more pronounced in the differential prediction case. The distribution of item difficulties also had a much larger effect in the differential prediction condition.

The most interesting effect was due to distribution of item difficulties. The uniform tests resulted in scores which were more fair to the minority subgroup than were scores on the peaked tests for almost all test lengths and degrees of item bias. The differences in *T*-fairness between the uniform and peaked tests were particularly large at the shortest and longest test lengths. At the highest level of item discrimination (a=1.1), the uniform tests showed a clear and substantial advantage over the peaked tests.

The differences that occurred between the uniform and peaked tests in the differential prediction condition were mainly due to the skewness and kurtosis of the predicted score distributions obtained in the respective conditions. As can be seen in Table 2, the uniform tests produced a predicted score distribution that was flatter and less skewed than that of the peaked tests. These differences in the shape of the predicted score distributions increased as item discrimination was increased.

The effect of the shape of the predicted score distribution is much greater in the differential prediction condition than in the majority prediction condition because of the relationships in the distribution between the mean of the score distributions and the selection cutoff. These effects can be seen in Figures 4 and 5. Figure 4 represents the case where majority prediction was used and the test items were biased against the minority subgroup. This situation will result in the mean $\hat{\theta}$ of the minority subgroup being below that of the majority subgroup. Since, in this case, such a small percentage of the minority subgroup is above the majority subgroup average, differences in the predicted distributions as a function of spread in item difficulties have a relatively small effect on *T*-fairness. However, with differential prediction (Figure 5) there will be no bias in the predicted means for either subgroup. Consequently, the effects of skewness and kurtosis on *T*-fairness are much larger.

When differential prediction was used, the uniform test was fairer to the minority subgroup than was the peaked test. This result was observed across test length and item discrimination conditions. For the higher discrimination levels, this result was consistent with the results from the validity data. Therefore, uniform tests are clearly preferable when used in combination with differential prediction. These results also imply that if differential predic-

tion is employed, it is possible to avoid the problem, often encountered using majority prediction, of trying to simultaneously minimize differential validity and *C*- or *T*-fairness.

SUMMARY AND CONCLUSIONS

This study was concerned with how test fairness, defined in terms of test validity and the models presented by Cleary and Thorndike, is influenced by test length, distribution of item difficulties, level of item discrimination and degree of item bias. The methodology involved computer simulation in which bias and fairness were represented in the context of latent trait theory. This approach eliminates many of the criterion measurement problems often present in empirical validation studies, and allows direct observation of the influence of item characteristics on test scores and on predictions made from those test scores. The situation assumed in the present study was that a single test was used to select an unrestricted sample of applicants from a hypothetical population consisting of a minority and a majority subgroup. The criterion on which the selections were validated was a unidimensional variable on which the subgroups had identical distributions.

Validity

The findings from the validity data indicated that contrary to the results of previous research, a uniform test often led to a higher validity for many practical test applications than did a peaked test In fact, if item discriminations were relatively high, uniform tests resulted in substantially higher validities than did peaked tests. More importantly, with respect to the issue of test fairness, the difference between subgroup validities could be reduced by using uniform rather than peaked tests. It was also found that validity differences such as those reported and often disputed in the testing literature, are to be expected when test items are biased against one of the applicant subgroups. The fact that such validity differences are not always found in empirical validation studies is probably due to the lack of power in the statistical tests used in these empirical investigations.

Selection Fairness Models

The shapes of both the subgroup score distributions and the predicted ability distributions were found to be very much affected by the characteristics of the items included in the selection instrument. Conclusions drawn from each of the models used for measuring selection fairness were a function of the predicted ability distributions. Consequently, selection fairness was found to be a function of a test's item characteristics as well.

Perhaps the most relevant finding for test construction was that certain combinations of item characteristics were more robust in the presence of item bias than were others. That is, item bias had less of an effect on fairness for some combinations of item discrimination, test lengths, and distribution of item difficulties, than for others. The relationships among these variables were very complex. In any practical application where it is necessary to know how a particular set of item characteristics will affect the fairness of a test, a simulation study should be implemented in which the conditions of the application are approximated as closely as possible.

Nevertheless, certain generalizations can be made based on the present results. If applicants are to be selected in a situation similar to the conditions assumed in this study, a test having a uniform spread of item difficulties will result in fairer predictions than will a peaked test, if a reasonably high level of item discrimination can be maintained. Also, the differential prediction model can be expected to provide fairer selection than will sole reliance on majority prediction equations. Furthermore, the advantages of using a uniform test will be enhanced in the differential prediction application.

The results from the differential prediction condition indicate the conservative nature of Cleary's fairness model as compared to Thorndike's model. The use of differential prediction results in tests that are perfectly fair according to the Cleary definition, yet substantial amounts of unfairness were indicated in terms of the Thorndike model. This is a phenomenon often reported in the literature on models of fairness; different models of fairness can sometimes lead to divergent implications about the fairness of a test in a given selection situation. Particularly when peaked tests are employed, these two fairness models will lead to different conclusions.

Future Research

The present study investigated only a limited class of test instruments. The conventional tests used are characterized by their use of an identical fixed sequence of items for all testees. Recently, a number of adaptive testing models have been developed as alternatives to the conventional model (see Weiss, 1974). In adaptive tests, items are selected on an individual basis for each testee. Research with adaptive tests (*e.g.*, McBride & Weiss, 1976; Vale & Weiss, 1975) has shown that they result in different score distributions than do conventional tests, with true ability held constant. Consequently, adaptive testing methods might result in different degrees of fairness in test scores. Future research should explore the fairness properties of adaptive testing models and compare them with those of conventional tests.

REFERENCES

Angoff, W. H. The investigation of test bias in the absence of an outside criterion. Paper presented at NIE Conference on Test Bias, Washington, DC, December 1975.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10(21), 95-115.

Atkin, R., Bray, R., Davison, M., Herzberger, S., Humphreys, L., & Selzer, U. Common factor differentiation, grades 5 through 11. Applied Psychological Measurement, 1977, 1(1) (In press).

Bartlett, C. J., Bobko, P., & Pine, S. M. Single-group validity: Fallacy of the facts? Journal of Applied Psychology (In press).

Bartlett, C. J., & O'Leary, B. S. A differential prediction model to moderate the effects of heterogeneous groups in personnel selection and classification. Personnel Psychology, 1969, 22, 1-17.

Betz, N. E., & Weiss, D. J. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

Breland, H. M., Stocking, M., Pinchak, B. M., & Abrams, N. The cross-cultural stability of mental test items: An investigation of response patterns for ten socio-cultural groups. (Research Report PR-74-2). Princeton, NJ: Educational Testing Service, February 1974.

Brogden, H. E. Variation in test validity with distribution of item difficulties, number of items, and degree of their intercorrelation. Psychometrika, 1946, 11, 197-214.

Campbell, J. T., Crooks, L. A., Mahoney, M. H., & Rock, D. A. An investigation of the sources of bias in the prediction of job performance: A six year study. (Research Report PR-73-37). Princeton, NJ: Educational Testing Service, 1973.

Cleary, T. A. Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5(2), 115-124.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J., & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952, 17, 127-147.

Echternacht, G. J. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Farr, L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, I. L., & Bartlett, C. J. Ethnic group membership as a moderator in prediction of job performance. (Technical Report #2). Washington, DC: American Institute for Research, September 1971.

Flaugher, R. L. Bias in testing: A review and discussion. (ERIC TM Report 36). Princeton, NJ: Educational Testing Service, May 1974.

Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority telephone operators. Journal of Applied Psychology, 1975, 60(4), 411-419.

Goldman, R. D., & Hewitt, B. N. An investigation of test bias for Mexican-American college students. Journal of Educational Measurement, 1975, 12(3), 187-196.

Goldman, R. D., & Richards, R. The SAT prediction of grades for Mexican-American versus Anglo-American students of the University of California, Riverside. Journal of Educational Measurement, 1974, 11, 129-135.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. Psychometrika, 1945, 10, 79-91.

Jensen, A. R. Test bias and construct validity. Revised address to the American Psychological Association Annual Meeting, Chicago, December 1975.

Jones, M. B. Moderated regression and equal opportunity. Educational and Psychological Measurement, 1973, 33, 591-602.

Kallingal, A. The prediction of grades for black and white students of Michigan State University. Journal of Educational Measurement, 1971, 8, 263-265.

Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 1952, 17, 181-193.

Lord, F. M. Formula scoring and number-right scoring. Journal of Educational Measurement, 1975, 12(1), 7-11.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy. (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

McNemar, Q. On so-called test bias. American Psychologist, 1975, 30(8), 848-851.

Petersen, N. S., & Novick, M. R. An evaluation of some models for test bias. (Technical Bulletin No. 23). Iowa City: The American College Testing Program, Research and Development Division, September 1974.

Pine, S. M., & Weiss, D. J. Psychometric issues in test bias and test fairness. Unpublished manuscript, University of Minnesota, Department of Psychology, Psychometric Methods Program, 1976.

Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 53, 5-9.

Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.

Standards for educational and psychological tests. Washington, DC: American Psychological Association, 1974.

Temp, G. Validity of the SAT for blacks and whites in thirteen integrated institutions. Journal of Educational Measurement, 1971, 8, 245-251.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Urry, V. W. A monte carlo study of logistic test models. Unpublished doctoral dissertation, Purdue University, 1969.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Washington, DC: Personnel Research and Development Center, U. S. Civil Service Commission, 1974 (In press).

Urry, V. W. Individual testing by Bayesian estimation. (Duplicated Report). Seattle: Bureau of Testing, University of Washington, 1971.

Vale, C. D., & Weiss, D. J. A simulation study of stradaptive ability testing. (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975. (NTIS No. AD A020961)

Weiss, D. J. Strategies of adaptive ability measurement. (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

APPENDIX

Table A
Means and Standard Deviations of Item Difficulty Distributions of Item Banks

	PEAKED TEST					
	Item Discrimination (a)					
Test	.3		.7		1.1	
Length	M	S.D.	M	S.D.	M	S.D.
10	.00	.09	.03	.12	-.07	.11
30	-.01	.08	.02	.11	-.02	.10
50	.00	.09	-.01	.11	.00	.10
70	.01	.09	.00	.11	.00	.10
100	.00	.09	.01	.11	.00	.10

	UNIFORM TEST					
	Item Discrimination (a)					
Test	.3		.7		1.1	
Length	M	S.D.	M	S.D.	M	S.D.
10	-.32	1.82	-.32	1.82	-.32	1.82
30	-.02	1.79	-.02	1.79	-.02	1.79
50	-.07	1.71	-.07	1.71	-.07	1.71
70	-.17	1.70	-.17	1.70	-.17	1.70
100	-.13	1.77	-.13	1.77	-.13	1.77

Table B
Score Distribution Characteristics for Conventional Tests of Length 10, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Mean		Standard Deviation				Skewness		Kurtosis	
					Uniform		Peaked					
a	Bias	Group	Uniform	Peaked	M.P.	D.P.	M.P.	D.P.	Uniform	Peaked	Uniform	Peaked
		True	-.074	-.074	1.006	1.006	1.006	1.006	-.01	-.01	.22	.22
.30												
		maj	-.074	-.074	.496	.496	.544	.544	-.23	-.14	-.22	-.24
	.5	min	-.198	-.198	.507	.495	.548	.546	-.24	-.05	-.08	-.26
	1	min	-.328	-.338	.507	.515	.554	.588	-.13	-.01	-.13	-.37
	2	min	-.583	-.604	.500	.526	.528	.543	.09	.30	-.16	-.28
.70												
		maj	-.074	-.074	.750	.750	.787	.787	-.27	-.17	-.32	-.70
	.5	min	-.357	-.393	.763	.749	.777	.801	-.21	.22	-.34	-.62
	1	min	-.875	-.697	.786	.769	.751	.806	-.01	.59	-.37	-.35
	2	min	-1.215	-1.216	.757	.777	.613	.761	.34	1.14	-.27	.88
1.1												
		maj	-.074	-.074	.825	.825	.874	.874	-.41	-.16	.12	-1.06
	.5	min	-.435	-.485	.880	.834	.877	.885	-.22	.32	-.30	-1.00
	1	min	-.813	-.854	.920	.849	.810	.858	-.14	.81	-.41	-.31
	2	min	-1.554	-1.372	.869	.829	.570	.758	.50	2.00	-.35	3.92

Note. M.P. is majority prediction equation; D.P. is differential prediction equation.

Table C
Score Distribution Characteristics for Conventional Tests of Length 30, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Mean		Standard Deviation				Skewness		Kurtosis	
					Uniform		Peaked					
a	Bias	Group	Uniform	Peaked	M.P.	D.P.	M.P.	D.P.	Uniform	Peaked	Uniform	Peaked
		True	-.074	-.074	1.006	1.006	1.006	1.006	-.01	-.01	.22	.22
.30												
		maj	-.074	-.074	.729	.729	.745	.745	.01	-.12	.16	-.18
	.5	min	-.332	-.329	.741	.746	.746	.759	.04	-.04	.14	-.05
	1	min	-.601	-.608	.738	.748	.745	.767	.04	.07	.04	.02
	2	min	-1.107	-1.097	.745	.754	.721	.764	.18	.28	.06	.24
.70												
		maj	-.074	-.074	.904	.904	.917	.917	-.12	-.13	-.05	-.61
	.5	min	-.451	-.508	.894	.904	.905	.923	.04	.23	-.15	-.61
	1	min	-.875	-.911	.896	.896	.873	.923	.22	.52	-.14	-.26
	2	min	-1.607	-1.598	.809	.885	.715	.866	.57	1.13	.23	1.10
.1												
		maj	-.074	-.074	.938	.938	.945	.945	-.15	-.12	.16	-1.10
	.5	min	-.507	-.526	.966	.942	.928	.947	.03	.34	-.07	-.95
	1	min	-.975	-.935	.976	.937	.847	.926	.12	.82	-.37	-.07
	2	min	-1.834	-1.544	.822	.560	.920	.868	.65	2.09	.06	4.79

Note. M.P. is majority prediction equation; D.P. is differential prediction equation.

Table D
Score Distribution Characteristics for Conventional Tests of Length 70, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Mean		Standard Deviation				Skewness		Kurtosis	
					Uniform		Peaked					
a	Bias	Group	Uniform	Peaked	M.P.	D.P.	M.P.	D.P.	Uniform	Peaked	Uniform	Peaked
		True	-.074	-.074	1.006	1.006	1.006	1.006	-.01	-.01	.22	.22
.30												
		maj	-.074	-.074	.851	.851	.853	.853	-.00	-.10	-.13	-.04
	.5	min	-.429	-.445	.876	.858	.864	.865	.03	.04	-.21	-.04
	1	min	-.781	-.810	.878	.860	.872	.865	.09	.11	-.16	-.07
	2	min	-1.488	-1.480	.882	.861	.835	.860	.22	.34	-.09	.03
.70												
		maj	-.074	-.074	.960	.960	.960	.960	-.15	-.11	-.25	-.67
	.5	min	-.508	-.538	.967	.959	.953	.961	-.02	.22	-.32	-.64
	1	min	-.977	-.978	.966	.954	.908	.954	.20	.53	-.29	-.29
	2	min	-1.828	-1.732	.973	.938	.719	.916	.52	1.16	-.03	1.17
1.1												
		maj	-.074	-.074	.979	.979	.967	.967	-.19	-.11	.04	-1.10
	.5	min	-.551	-.541	1.003	.977	.952	.963	-.15	.37	-.32	-.90
	1	min	-1.048	-.973	.988	.857	.857	.942	.20	.88	-.42	-.02
	2	min	-1.945	-1.595	.879	.955	.551	.842	.68	2.13	.16	4.86

Note. M.P. is majority prediction equation; D.P. is differential prediction equation.

Table E
Score Distribution Characteristics for Conventional Tests of Length 100, as a Function of Discrimination (a), Bias, and Group, for Uniform and Peaked Tests

			Mean		Standard Deviation				Skewness		Kurtosis	
					Uniform		Peaked					
a	Bias	Group	Uniform	Peaked	M.P.	D.P.	M.P.	D.P.	Uniform	Peaked	Uniform	Peaked
		True	-.074	-.074	1.006	1.006	1.006	1.006	-.01	-.01	.22	.22
.30												
		maj	-.074	-.074	.889	.889	.893	.893	-.06	0.16	-.05	.07
	.5	min	-.456	-.479	.903	.892	.918	.902	-.00	-.05	-.13	-.02
	1	min	-.837	-.880	.904	.893	.914	.898	.05	.08	-.04	-.05
	2	min	-1.606	-1.638	.911	.891	.875	.895	.22	.28	-.09	-.03
.70												
		maj	-.074	-.074	.972	.972	.972	.972	-.14	-.13	-.24	.65
	.5	min	-.526	-.553	.971	.971	.969	.973	.03	.19	-.24	-.64
	1	min	-.993	-1.011	.962	.968	.922	.965	.19	.52	-.22	-.28
	2	min	-1.871	-1.782	.866	.955	.727	.930	.53	1.14	.00	1.08
1.1												
		maj	-.074	-.074	.987	.987	.972	.972	-.10	-.12	-.10	-1.07
	.5	min	-.554	-.548	1.004	.985	.960	.968	.06	.37	-.23	-.90
	1	min	-1.049	-.985	.981	.981	.863	.947	.19	.88	-.36	-.05
	2	min	-1.958	-1.616	.875	.966	.554	.846	.63	2.16	.14	5.03

Note: M.P. is majority prediction equation; D.P. is differential prediction equation.

Table F
C-Index for Uniform (U) and Peaked (P) Tests

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.5	min	-.124	-.124	-.258	-.255	-.317	-.328	-.355	-.371	-.382	-.405
		diff	-.124	-.124	-.258	-.255	-.317	-.328	-.355	-.371	-.382	-.405
	1.0	min	-.254	-.264	-.527	-.534	-.635	-.664	-.707	-.736	-.763	-.806
		diff	-.254	-.264	-.527	-.534	-.635	-.664	-.707	-.736	-.763	-.806
	2.0	min	-.509	-.530	-1.033	-1.023	-1.262	-1.286	-1.414	-1.406	-1.532	-1.564
		diff	-.509	-.530	-1.033	-1.023	-1.262	-1.286	-1.414	-1.406	-1.532	-1.564
.70												
	0.0	maj	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.5	min	-.283	-.319	-.377	-.434	-.422	-.461	-.434	-.464	-.452	-.479
		diff	-.283	-.319	-.377	-.434	-.422	-.461	-.434	-.464	-.452	-.479
	1.0	min	-.586	-.623	-.801	-.837	-.879	-.894	-.903	-.904	-.919	-.937
		diff	-.586	-.623	-.801	-.837	-.879	-.894	-.903	-.904	-.919	-.937
	2.0	min	-1.141	-1.142	-1.533	-1.524	-1.675	-1.634	-1.754	-1.658	-1.797	-1.708
		diff	-1.141	-1.142	-1.533	-1.524	-1.675	-1.634	-1.754	-1.658	-1.797	-1.708
1.1												
	0.0	maj	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.5	min	-.361	-.411	-.433	-.452	-.462	-.462	-.477	-.467	-.480	-.474
		diff	-.361	-.411	-.433	-.452	-.462	-.462	-.477	-.467	-.480	-.474
	1.0	min	-.739	-.780	-.901	-.861	-.946	-.882	-.974	-.899	-.975	-.911
		diff	-.739	-.780	-.901	-.861	-.946	-.882	-.974	-.899	-.975	-.911
	2.0	min	-1.480	-1.298	-1.760	-1.470	-1.778	-1.499	-1.871	-1.521	-1.884	-1.542
		diff	-1.480	-1.298	-1.760	-1.470	-1.778	-1.499	-1.871	-1.521	-1.884	-1.542

Table G
T-Index for Uniform (U) and Peaked (P) Tests, Using Majority Prediction

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	38.4	56.8	45.4	47.4	41.6	43.8	44.0	46.2	44.0	49.8
	.5	min	30.4	46.2	31.8	33.8	28.0	28.2	29.6	30.2	30.6	30.8
		diff	-8.0	-10.6	-13.6	-13.6	-13.6	-15.6	-14.4	-16.0	-13.4	-19.0
	1.0	min	21.6	37.4	20.4	21.4	18.0	17.2	18.2	19.0	17.2	19.2
		diff	-16.8	-19.4	-25.0	-26.0	-23.6	-26.6	-25.8	-27.2	-26.8	-30.6
	2.0	min	10.0	20.2	7.8	7.2	4.4	3.6	4.4	4.6	4.6	4.6
		diff	-28.4	-36.6	-37.6	-40.2	-37.2	-40.2	-39.6	-41.6	-39.4	-45.2
.70												
	0.0	maj	40.6	53.0	50.2	45.0	46.2	49.4	48.0	49.8	48.2	49.2
	.5	min	26.6	35.4	34.4	29.2	28.4	30.4	32.0	29.8	31.0	29.2
		diff	-14.0	-17.6	-15.8	-15.8	-17.8	-19.0	-16.0	-20.0	-17.2	-20.0
	1.0	min	16.4	22.8	19.6	15.0	14.8	15.6	16.2	15.4	15.4	14.8
		diff	-24.2	-30.2	-30.6	-30.0	-31.4	-33.8	-31.8	-34.4	-32.8	-34.4
	2.0	min	4.6	6.2	3.8	3.6	2.8	3.6	2.4	3.4	3.0	3.2
		diff	-36.0	-46.8	-46.4	-41.4	-43.4	-45.8	-45.6	-46.4	-45.2	-46.0
1.1												
	0.0	maj	40.0	52.6	47.2	46.8	47.8	47.6	47.8	48.4	48.8	50.0
	.5	min	25.2	35.2	29.6	29.4	28.0	29.2	27.8	29.2	29.0	29.2
		diff	-14.8	-17.4	-17.6	-17.4	-19.8	-18.4	-20.0	-19.2	-19.8	-20.8
	1.0	min	15.0	20.4	17.4	15.2	15.8	14.6	13.8	15.4	15.4	15.4
		diff	-25.0	-32.2	-29.8	-31.6	-32.0	-33.0	-34.0	-33.0	-33.4	-34.6
	2.0	min	3.4	4.8	3.0	3.4	2.4	3.4	2.2	3.0	2.6	2.6
		diff	-36.6	-47.8	-44.2	-43.4	-45.4	-44.2	-45.6	-45.4	-46.2	-47.4

Table H
T-Index for Uniform (U) and Peaked (P) Tests, Using Differential Prediction

			Test Length									
			10		30		50		70		100	
a	Bias	Group	U	P	U	P	U	P	U	P	U	P
.30												
	0.0	maj	38.4	56.8	45.4	47.4	41.6	43.8	44.0	46.2	44.0	49.8
	.5	min	52.6	46.2	40.6	42.6	48.4	47.8	45.0	42.0	46.8	45.4
		diff	14.2	-10.6	-4.8	-4.8	6.8	4.0	1.0	-4.2	2.8	-4.4
	1.0	min	41.8	37.4	49.2	47.8	47.2	44.4	44.4	46.2	46.0	44.2
		diff	3.4	-19.4	3.8	.4	5.6	.6	.4	0.0	2.0	-5.6
	2.0	min	43.6	33.0	41.4	39.0	44.6	45.2	41.4	44.2	44.0	42.4
		diff	5.2	-23.8	-4.0	-8.4	3.0	1.4	-2.6	-2.0	0.0	-7.4
.70												
	0.0	maj	40.6	53.0	50.2	45.0	46.2	49.4	48.0	49.8	48.2	49.2
	.5	min	47.8	47.8	48.2	41.6	47.0	45.8	47.4	46.4	45.8	46.4
		diff	7.2	-5.2	-2.0	-3.4	.8	-3.6	-.6	-3.4	-2.4	-2.8
	1.0	min	49.8	46.0	45.0	41.4	47.0	44.8	45.6	41.8	47.6	42.8
		diff	9.2	-7.0	-5.2	-3.6	.8	-4.6	-2.4	-8.0	-.6	-6.4
	2.0	min	41.2	31.2	44.0	36.6	40.8	38.2	39.4	38.8	42.4	37.4
		diff	.6	-21.8	-6.2	-8.4	-5.4	-11.2	-8.6	-11.0	-5.8	-11.8
1.1												
	0.0	maj	40.0	52.6	47.2	46.8	47.8	47.6	47.8	48.4	48.8	50.0
	.5	min	46.0	43.6	43.8	42.0	45.6	44.0	48.2	43.2	47.4	45.0
		diff	6.0	-9.0	-3.4	-4.8	-2.2	-3.6	.4	-5.2	-1.4	-5.0
	1.0	min	48.4	36.2	48.8	39.6	42.6	39.6	46.6	38.4	45.6	39.4
		diff	8.4	-16.4	1.6	-7.2	-5.2	-8.0	-1.2	-10.0	-3.2	-10.6
	2.0	min	51.0	27.6	40.8	28.6	43.0	30.8	43.0	30.8	44.0	29.6
		diff	11.0	-25.0	-6.4	-18.2	-4.8	-16.8	-4.8	-17.6	-4.8	-20.4

10

Previous Reports in this Series

73-1. Weiss, D.J. & Betz, N.E. Ability Measurement: Conventional or Adaptive? February 1973. (AD 757788).

73-2. Bejar, I.I. & Weiss, D.J. Comparison of Four Empirical Item Scoring Procedures. August 1973.

73-3. Weiss, D.J. The Stratified Adaptive Computerized Ability Test. September 1973. (AD 768376).

73-4. Betz, N.E. & Weiss, D.J. An Empirical Study of Computer-Administered Two-Stage Ability Testing. October 1973. (AD 768993).

74-1. DeWitt, L.J. & Weiss, D.J. A Computer Software System for Adaptive Ability Measurement. January 1974. (AD 773961).

74-2. McBride, J.R. & Weiss, D.J. A Word Knowledge Item Pool for Adaptive Ability Measurement. June 1974. (AD 781894).

74-3. Larkin, K.C. & Weiss, D.J. An Empirical Investigation of Computer-Administered Pyramidal Ability Testing. July 1974. (AD 783553).

74-4. Betz, N.E. & Weiss, D.J. Simulation Studies of Two-Stage Ability Testing. October 1974. (AD A001230).

74-5. Weiss, D.J. Strategies of Adaptive Ability Measurement. December 1974. (AD A004270).

75-1. Larkin, K.C. & Weiss, D.J. An Empirical Comparison of Two-Stage and Pyramidal Adaptive Ability Testing. February 1975. (AD A006733).

75-2. McBride, J.R. & Weiss, D.J. TETREST: A FORTRAN IV Program for Calculating Tetrachoric Correlations. March 1975. (AD A007572).

75-3. Betz, N.E. & Weiss, D.J. Empirical and Simulation Studies of Flexilevel Ability Testing. July 1975. (AD A013185).

75-4. Vale, C.D. & Weiss, D.J. A Study of Computer-Administered Stradaptive Ability Testing. October 1975. (AD A018758).

75-5. Weiss, D.J. (Ed.). Computerized Adaptive Trait Measurement: Problems and Prospects. November 1975. (AD A018675).

75-6. Vale, C.D. & Weiss, D.J. A Simulation Study of Stradaptive Ability Testing. December 1975. (AD A020961).

76-1. McBride, J.R. & Weiss, D.J. Some Properties of a Bayesian Adaptive Ability Testing Strategy. March 1976. (AD A022964).

76-2. Miller, T.W. & Weiss, D.J. Effects of Time-Limits on Test-Taking Behavior. April 1976. (AD A024422).

76-3. Betz, N.E. & Weiss, D.J. Effects of Immediate Knowledge of Results and Adaptive Testing on Ability Test Performance. June 1976. (AD A027147).

76-4. Betz, N.E. & Weiss, D.J. Psychological Effects of Immediate Knowledge of Results and Adaptive Ability Testing. June 1976. (AD A027170).

Weiss, D.J. Final Report: Computerized Ability Testing, 1972-1975. April 1976. (AD A024516).

AD Numbers are those assigned by the Defense Documentation Center, for retrieval through the National Technical Information Service.

Copies of these reports are available, while supplies last, from:

Psychometric Methods Program
Department of Psychology
University of Minnesota
Minneapolis, Minnesota 55455

35